Topic 10: Simple Linear Regression

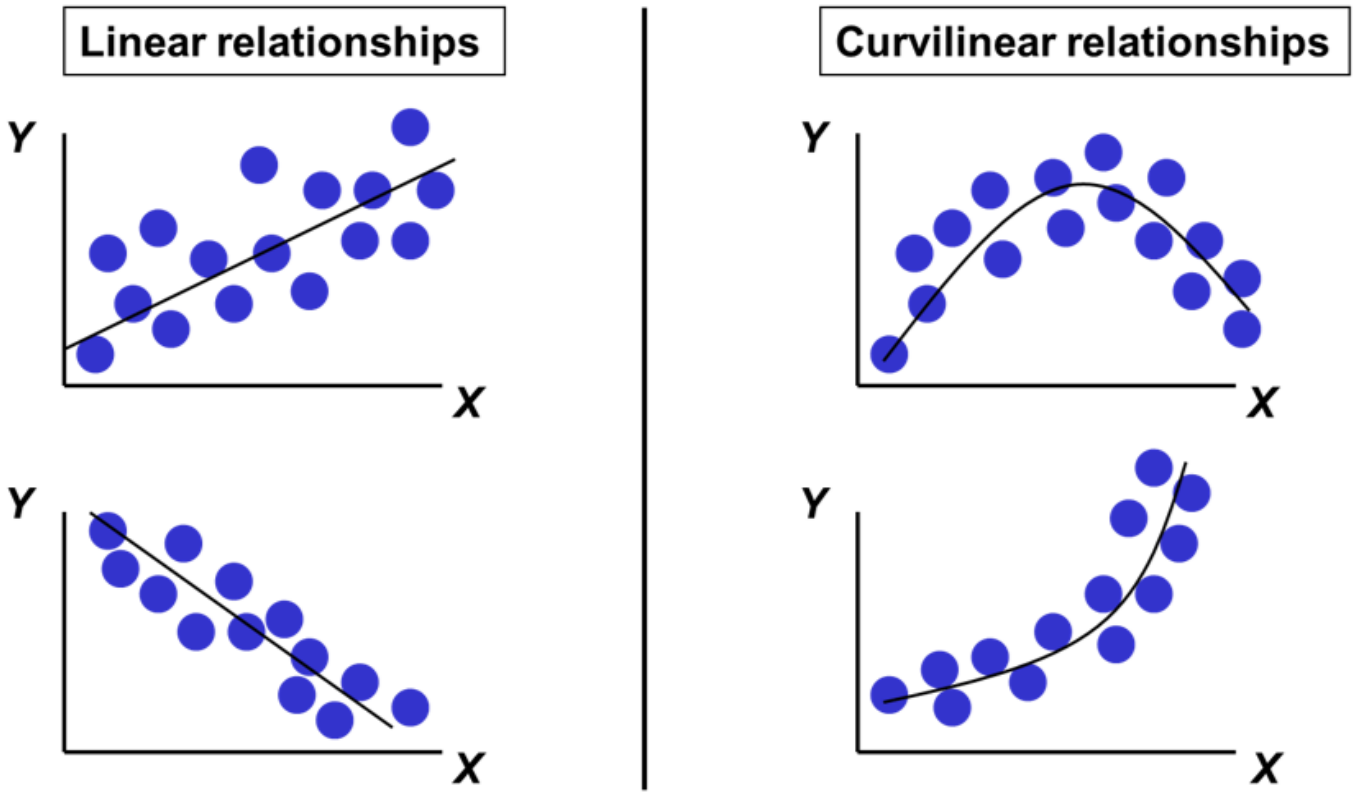
Dr. Lei Pan

School of Accounting, Economics and Finance
Curtin University

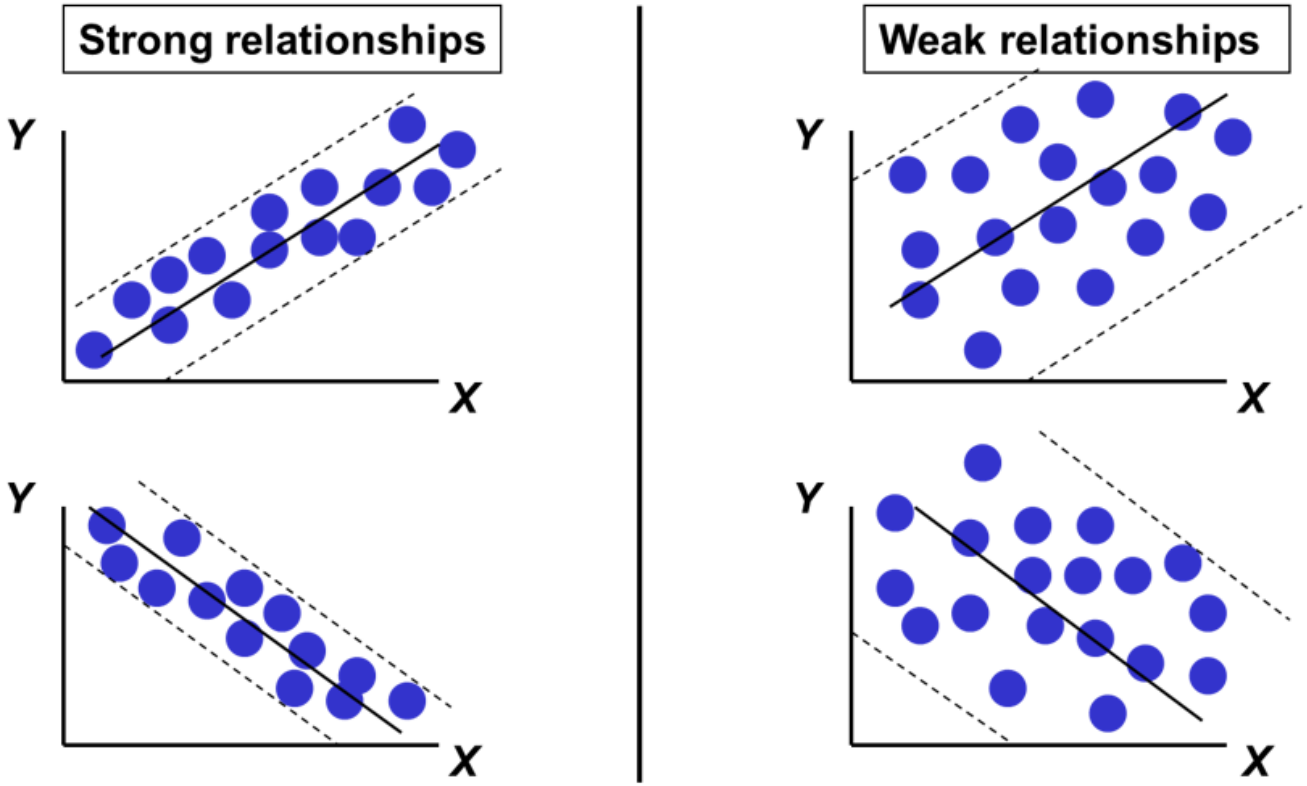
Correlation vs. Regression

- A scatter plot can be used to show the relationship between two variables.
- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables.
  - Correlation is only concerned with strength of the relationship.
  - No causal effect is implied with correlation.

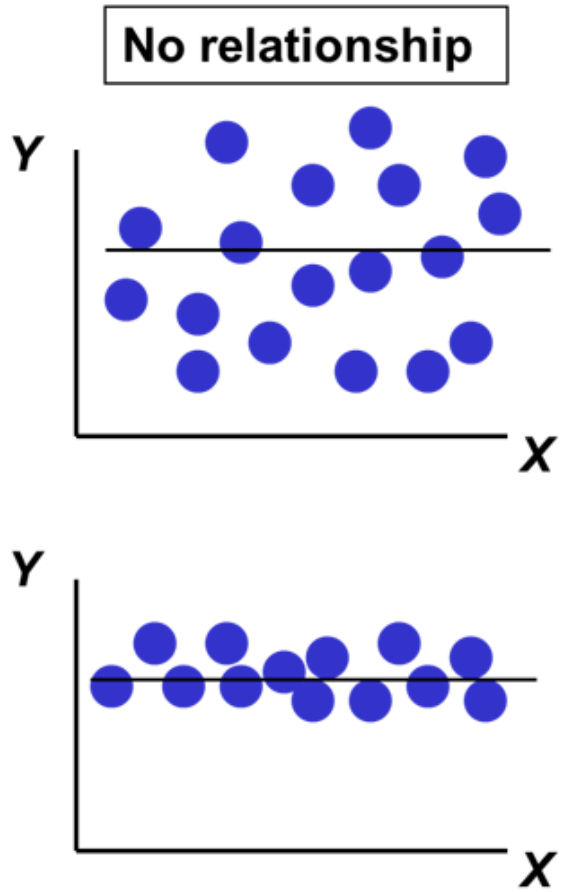
Types of Relationships (1 of 3)



Types of Relationships (2 of 3)



Types of Relationships (3 of 3)



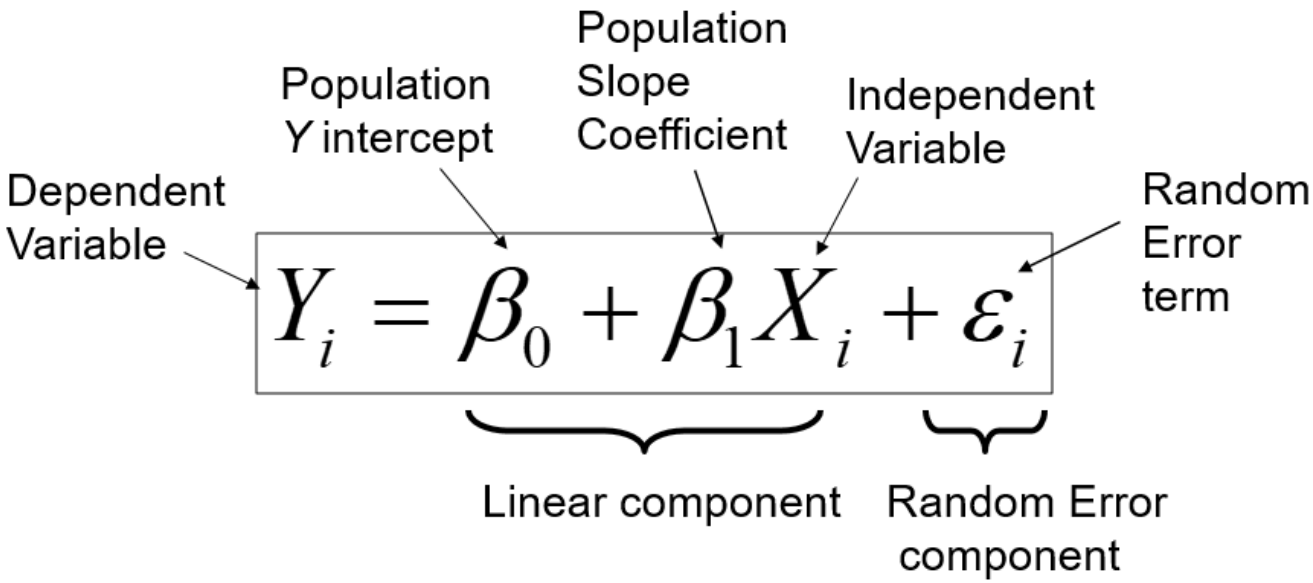
Introduction to Regression Analysis

- **Regression analysis** is used to:
  - Predict the value of a dependent variable (Y) based on the value of at least one independent variable (X).
  - Explain the impact of changes in an independent variable on the dependent variable.
- **Dependent variable** (**Y**): the variable we wish to predict or explain (response variable)
- **Independent variable** (**X**): the variable used to explain the dependent variable (explanatory variable)

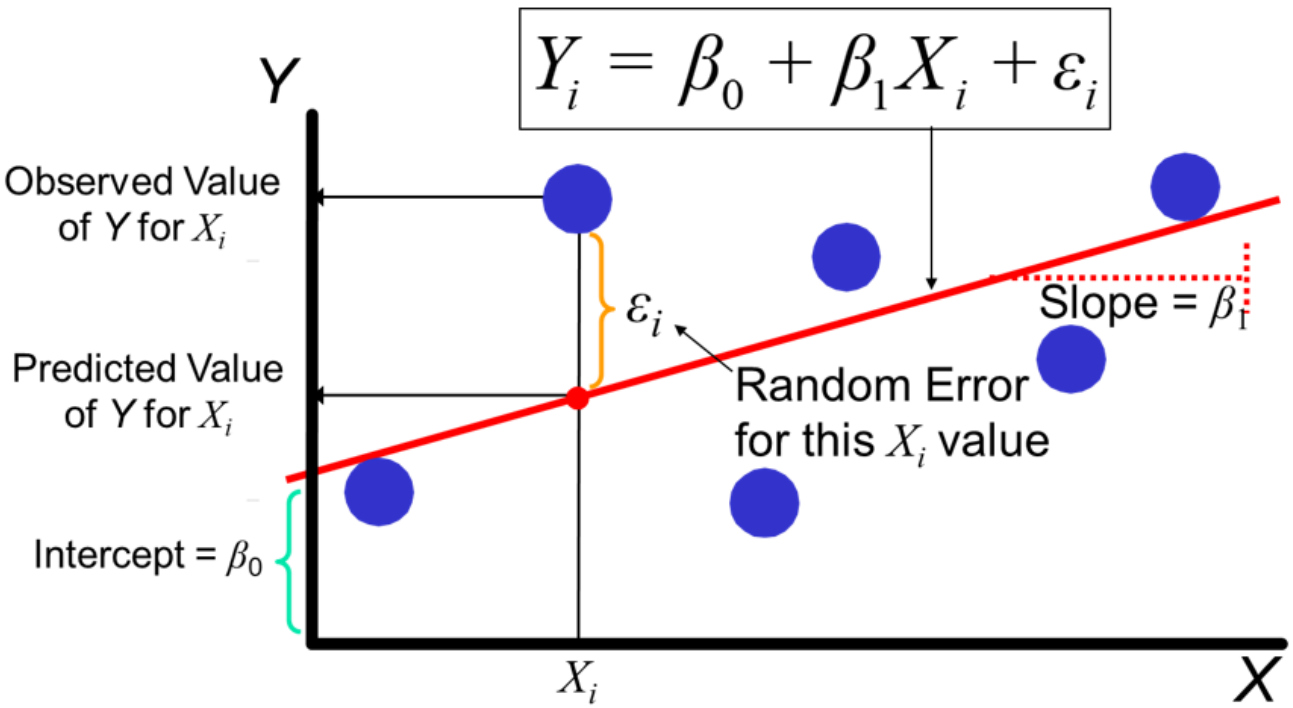
Simple Linear Regression Model (1 of 3)

- **Simple linear regression**:
  - Only **one** independent variable (X).
  - Relationship between X and Y is described by a linear function.
  - Changes in Y are assumed to be caused by changes in X.

Simple Linear Regression Model (2 of 3)

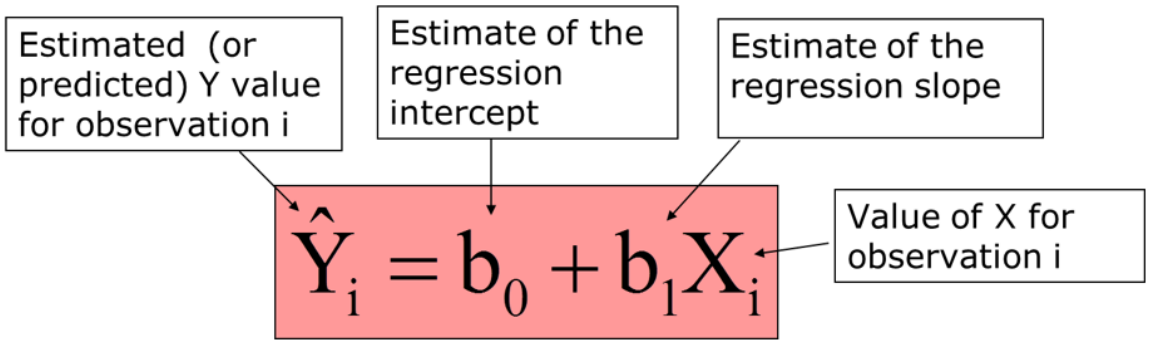


Simple Linear Regression Model (3 of 3)



Simple Linear Regression Equation (Prediction Line)

- The simple linear regression equation provides an estimate of the population regression line.



- The individual random error terms e_i have a mean of zero.

The Least Square Method

- b_0 and b_1 are obtained by finding the values that minimise the sum of the squared differences between actual values (Y) and predicted values ($\hat{Y}$)

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

- The coefficients b_0 and b_1, and other regression results in this topic, will be found using Excel.

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of Y when the value of X is zero
- b_1 is the estimated change in the average value of Y as a result of a one-unit change in X

Example: Simple Linear Regression

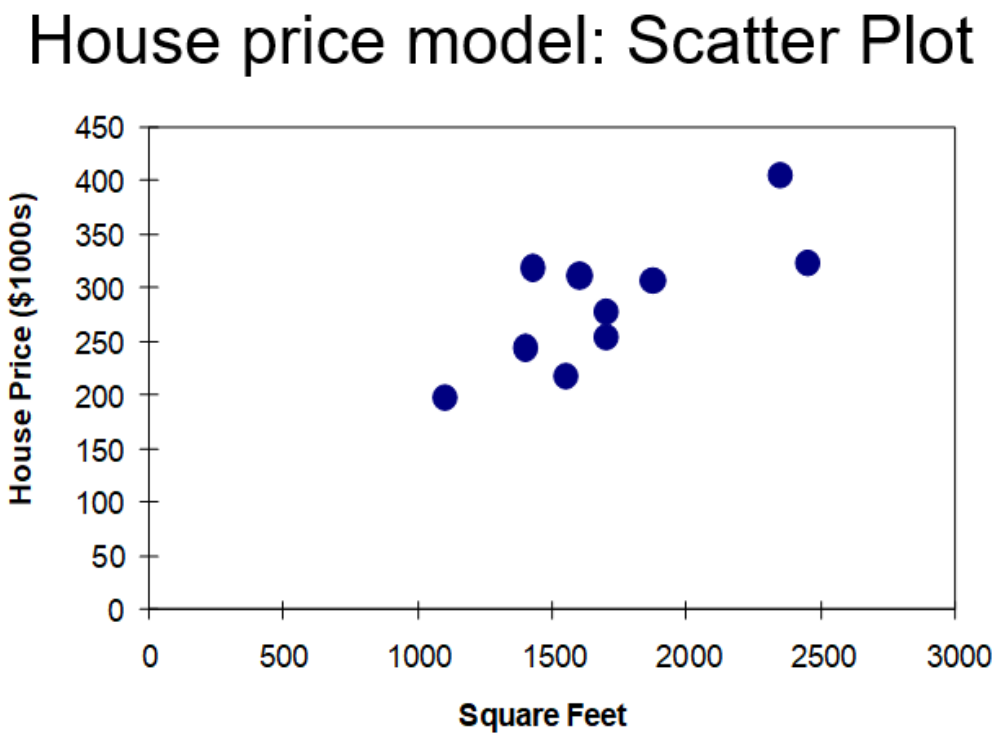
- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet).
- A random sample of 10 houses is selected.
  - Dependent variable (Y) = house price in \$1000s
  - Independent variable (X) = square feet



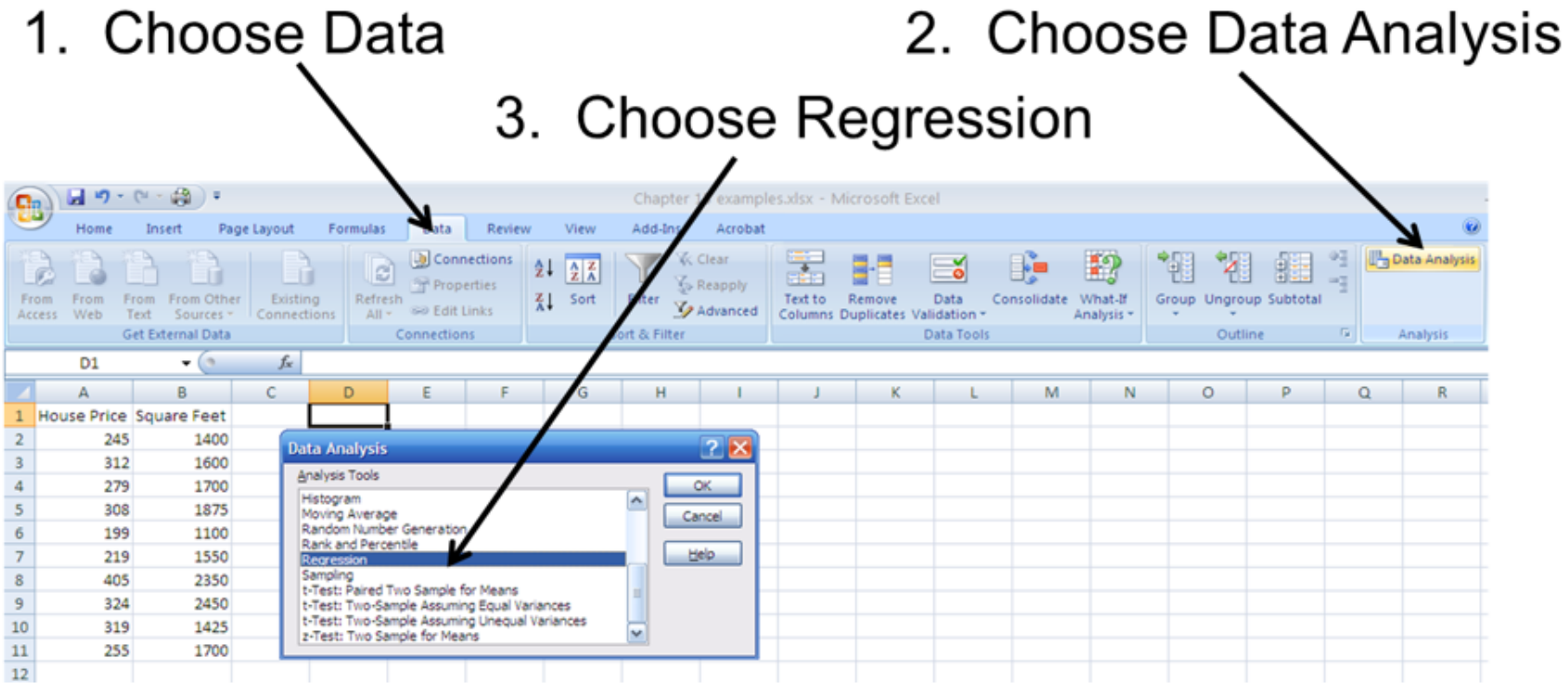
Example: Data

House Price in $1000s (*Y*)	Square Feet (*X*)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Example: Scatter Plot

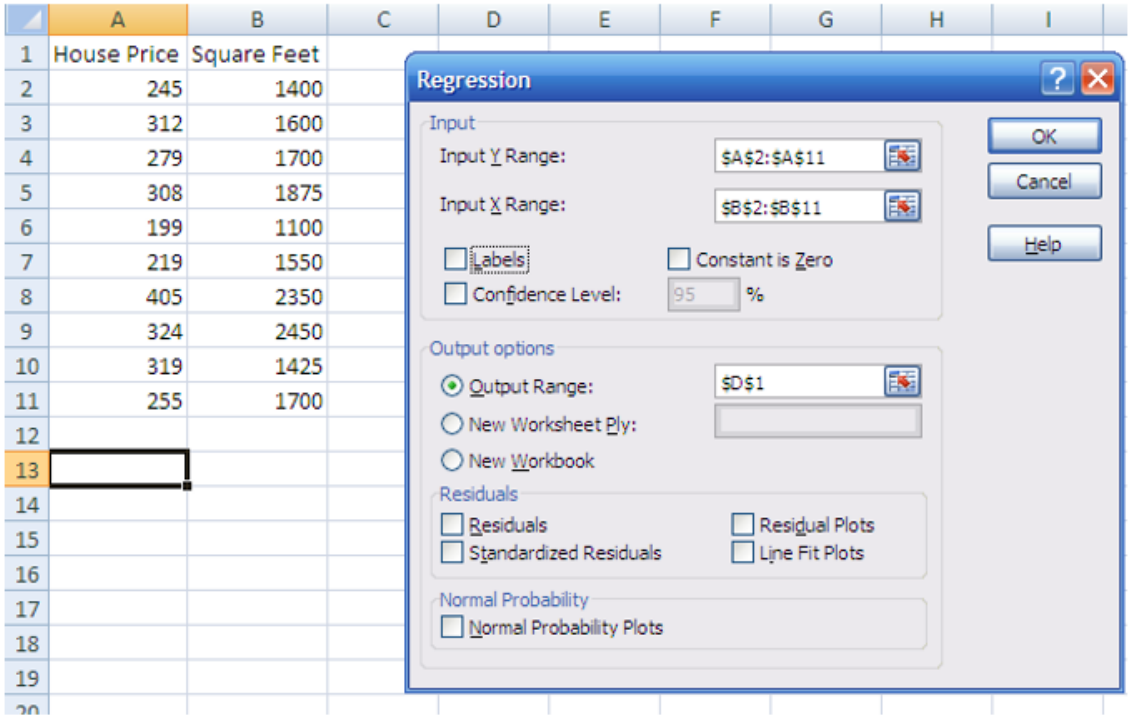


Example: Using Data Analysis ToolPak (1 of 2)



Example: Using Data Analysis ToolPak (2 of 2)

Enter Y range and X range and desired options.



Example: Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

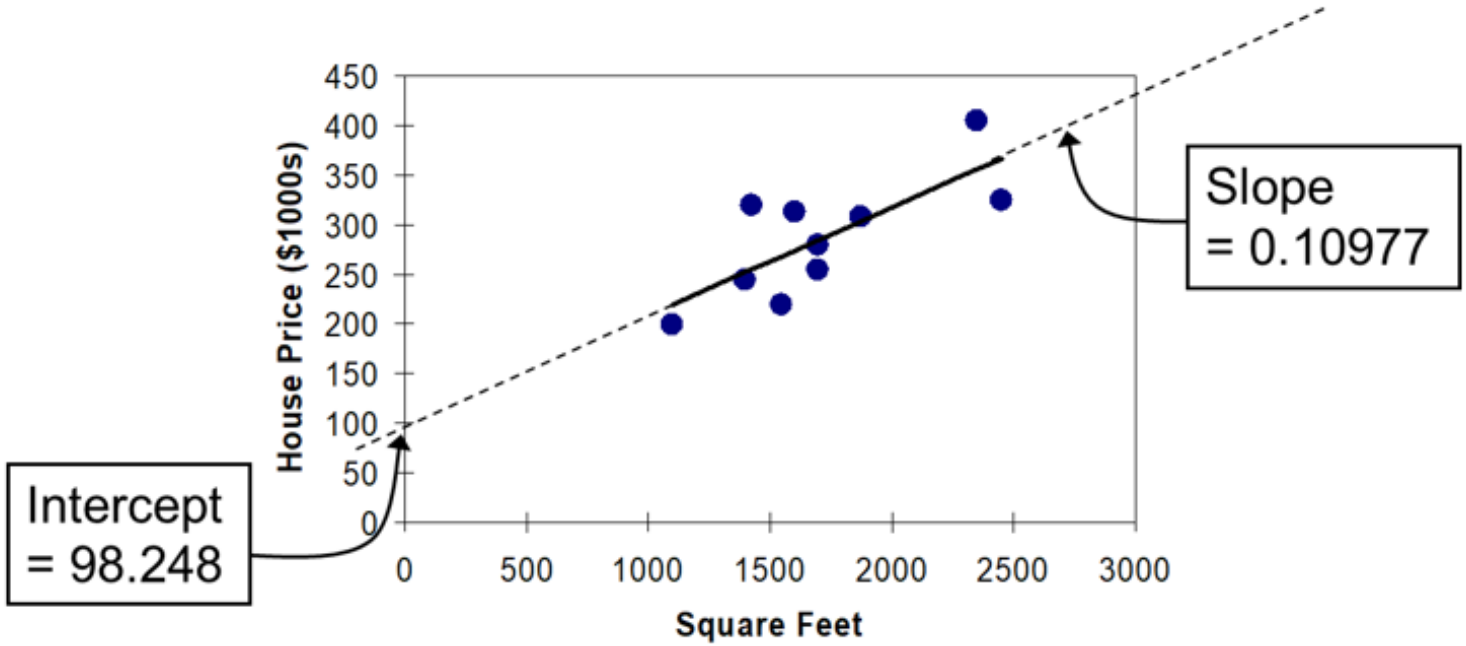
$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Example: Graphical Representation

House price model: Scatter Plot and Prediction Line



Example: Interpretation of b_0

$$\hat{\text{House}}\text{ price} = \boxed{98.24833} + 0.10977(\text{square feet})$$

- b_0 is the estimated mean value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)
- Because a house cannot have a square footage of 0, b_0 has no practical application.

Example: Interpretation of b_1

$$\hat{\text{House}} \text{ price} = 98.24833 + \boxed{0.10977} \text{ (square feet)}$$

- b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X
- Here, $b_1 = 0.10977$ tells us that the mean value of a house increases by $0.10977 \times \$1,000 = \$109.77$, on average, for each additional one square foot of size.

Example: Making Predictions (1 of 2)

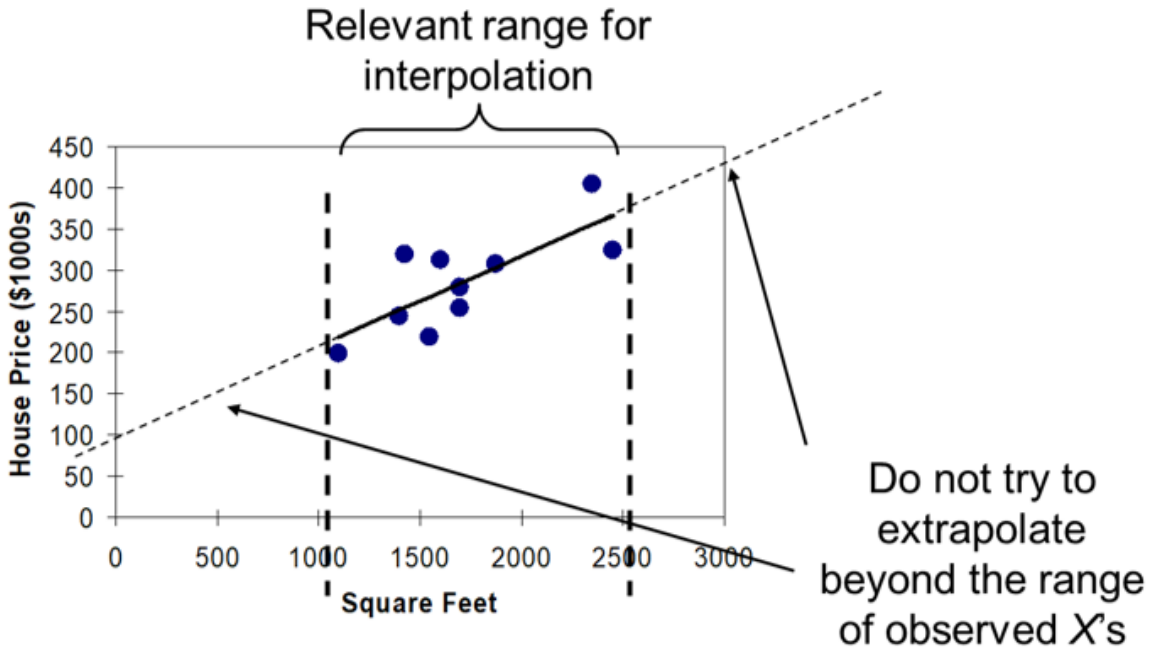
Predict the price for a house with 2000 square feet:

$$\begin{aligned}\widehat{\text{House price}} &= 98.25 + 0.1098\ (\text{sq.ft.}) \\ &= 98.25 + 0.1098 \times 2000 \\ &= 317.85\end{aligned}$$

The predicted price for a house with 2000 square feet is 317.85 ($1,000s) = $317,850

Example: Making Predictions (2 of 2)

- When using a regression model for prediction, only predict within the relevant range of data.



You Do It

A regression between foot length (response variable in cm) and height (explanatory variable in inches) for 33 students resulted in the following regression equation:

$$\hat{y} = 10.9 + 0.23x$$

(a) One student in the sample was 73 inches tall with a foot length of 29 cm. What is the predicted foot length for this student?

(b) One student in the sample was 73 inches tall with a foot length of 29 cm. What is the residual for this student?

(c) What is the estimated average foot length for students who are 70 inches tall?

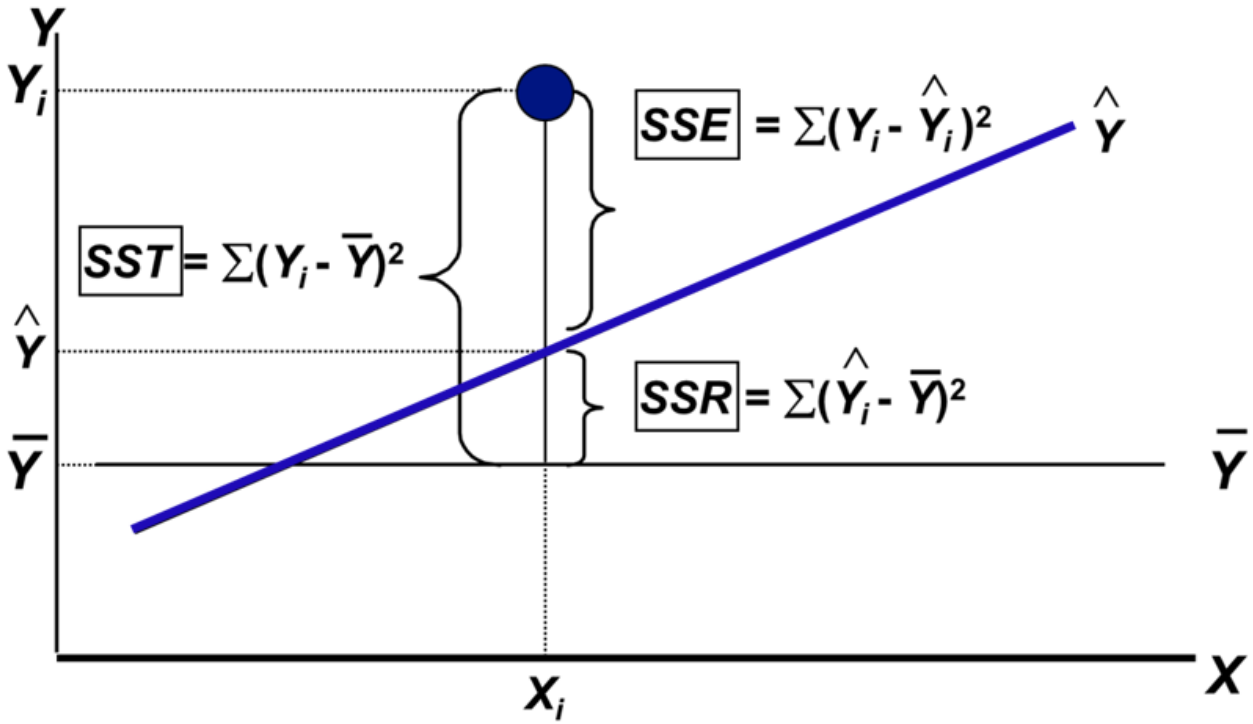
Measures of Variation (1 of 2)

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

SST: Total sum of squares (Total Variation)	**SSR: Regression sum of squares (Explained Variation)**	**SSE: Error sum of squares (Unexplained Variation)**
$\text{SST} = \sum(\text{Y}_\text{i} - \overline{\text{Y}})^2$	$SSR = \sum(\hat{Y}_i - \overline{Y})^2$	$SSE = \sum(Y_i - \hat{Y}_i)^2$
Measures the variation of the Y_i values around their mean Y	Explained variation attributable to the relationship between X and Y	Variation attributable to factors other than the relationship between X and Y

Measures of Variation (2 of 2)



Coefficient of Determination, r^2

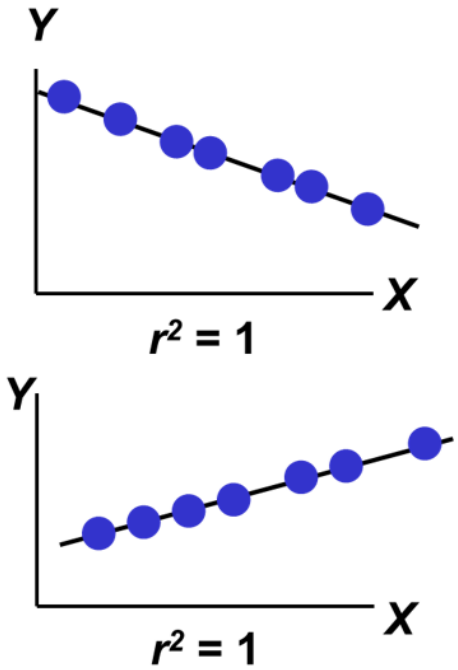
- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.
- The coefficient of determination is also called **r-squared** and is denoted as r^2.

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Note: $0 \leq r^2 \leq 1$

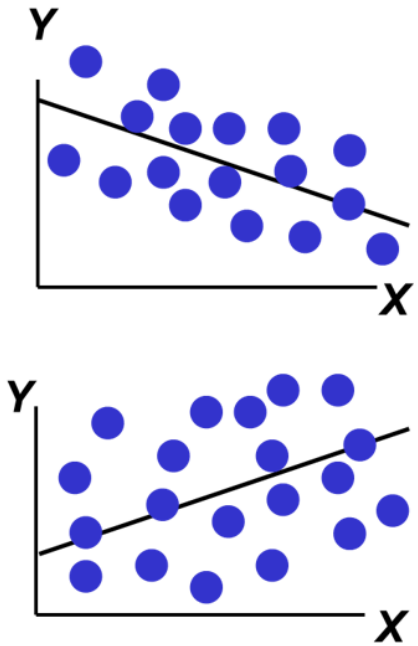
Example: Approximate r-squared Values (1 of 3)

- Perfect linear relationship between X and Y:
  - 100% of the variation in Y is explained by variation in X



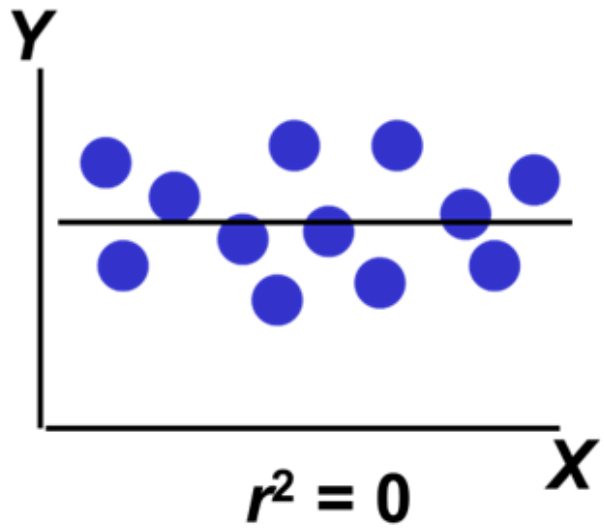
Example: Approximate r-squared Values (2 of 3)

- Weaker linear relationships between X and Y:
  - Some but not all of the variation in Y is explained by variation in X.
  - $0 < r^2 < 1$



Example: Approximate r-squared Values (3 of 3)

- No linear relationship between X and Y:
  - The value of Y does not depend on X (None of the variation in Y is explained by variation in X).
  - $r^2 = 0$



Example: Coefficient of Determination, r^2 in Excel

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Curtin University

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

where

SSE = error sum of squares

n = sample size

Example: Standard Error of Estimate in Excel

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$S_{YX} = 41.33032$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

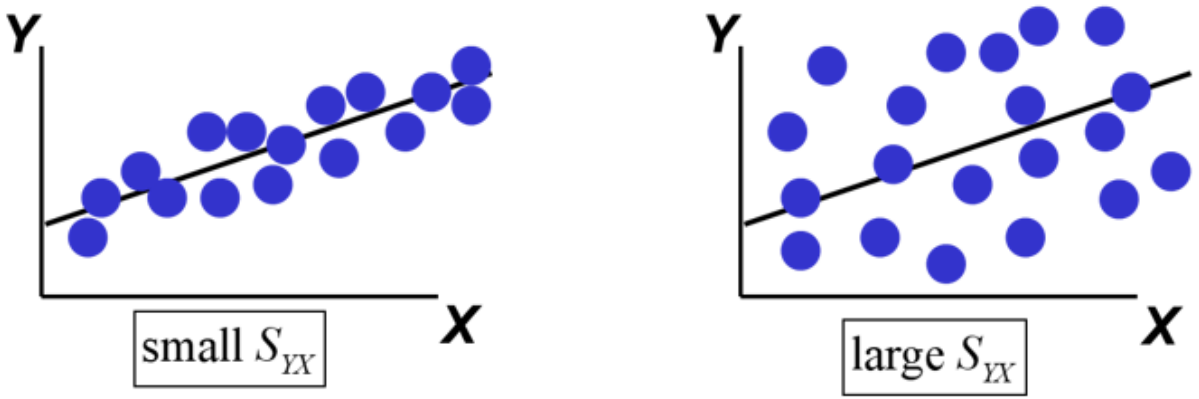
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Curtin University

50 YEARS OF INNOVATION

Comparing Standard Errors

- S_{YX} is a measure of the variation of observed Y values from the regression line.



- The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data.
  - i.e. $S_{YX} = \$41.33K$ is moderately small relative to house prices in the \$200K - \$400K range

Assumptions of Regression

Use the acronym LINE:

- Linearity
  - The relationship between X and Y is linear.
- Independence of Errors
  - Error values are statistically independent.
  - Particularly important when data are collected over a period of time.
- Normality of Error
  - Error values are normally distributed for any given value of X.
- Equal Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance.

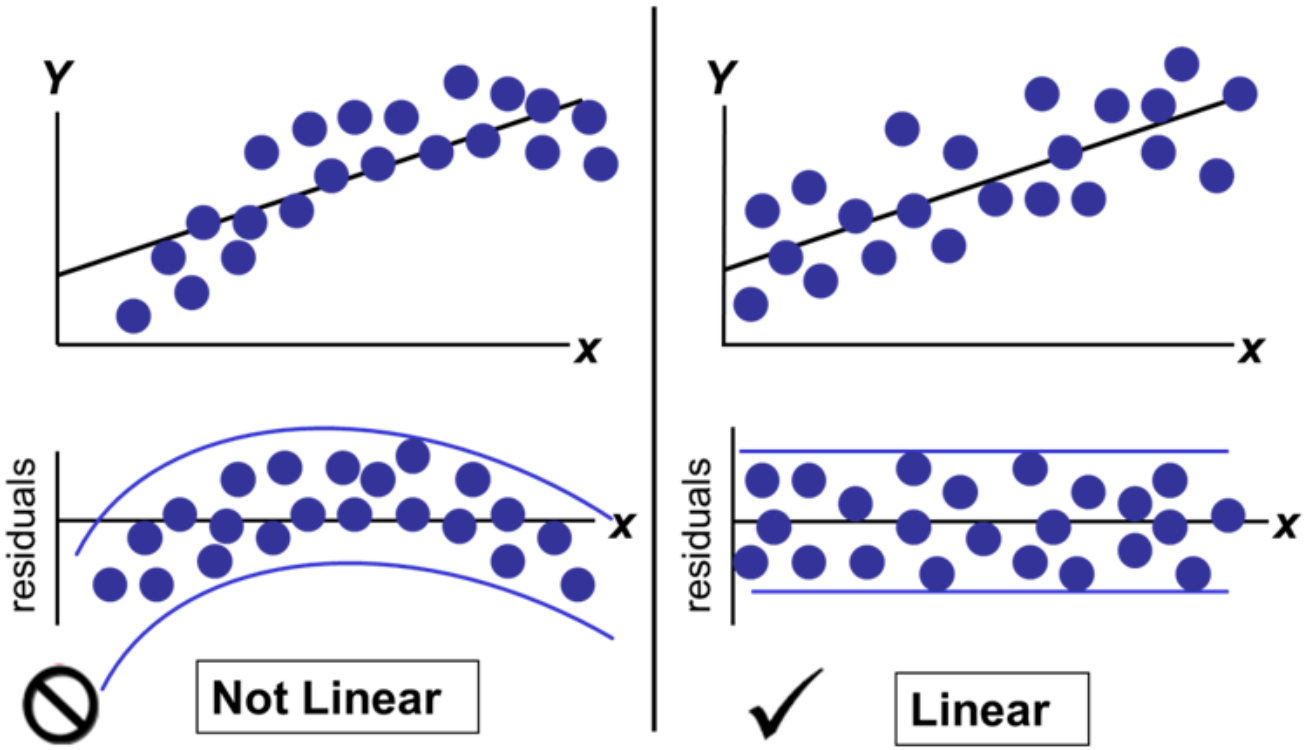
Residual Analysis

- The residual for observation i, e_i, is the difference between its observed and predicted value.

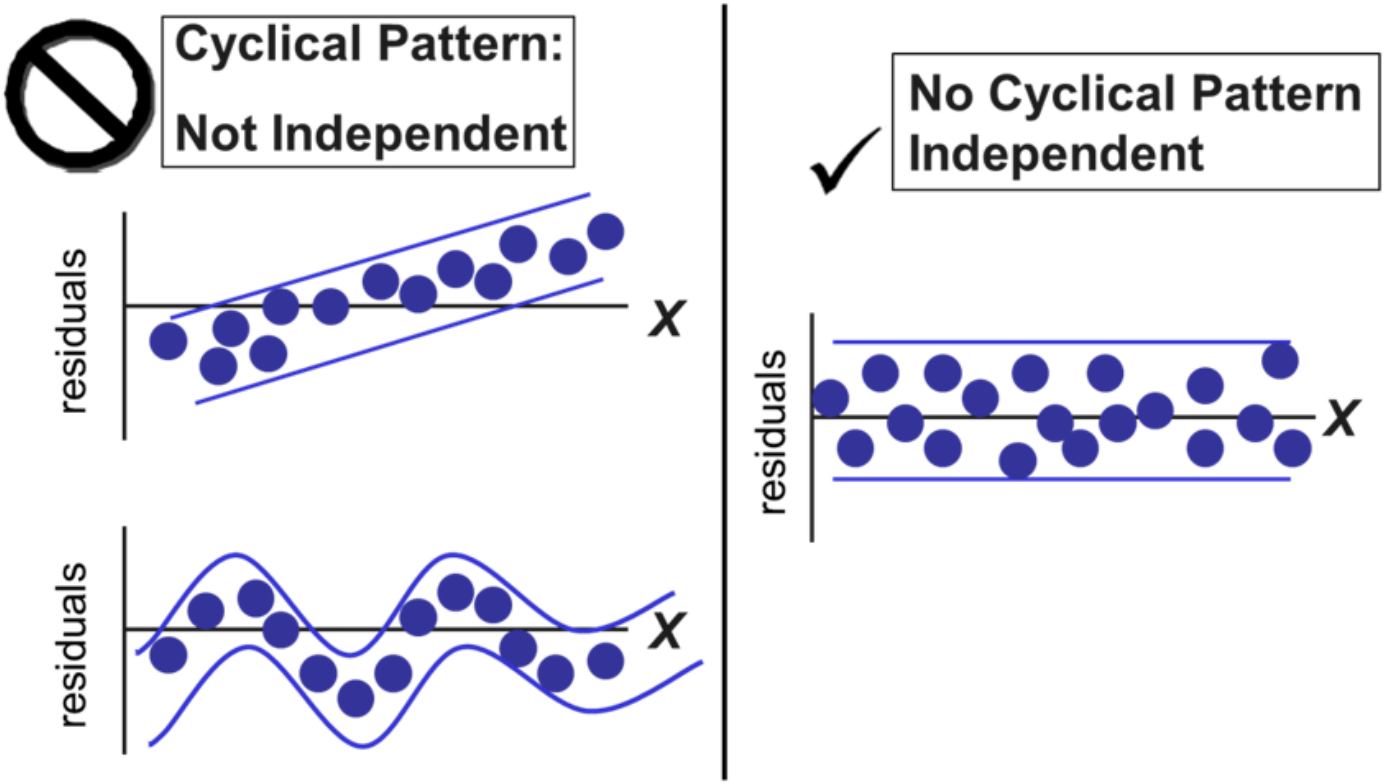
$$e_i = Y_i - \hat{Y}_i$$

- Check the assumptions of regression by examining the residuals.
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
- Graphical analysis of residuals
  - Can plot residuals vs. X

Residual Analysis for Linearity



Residual Analysis for Independence

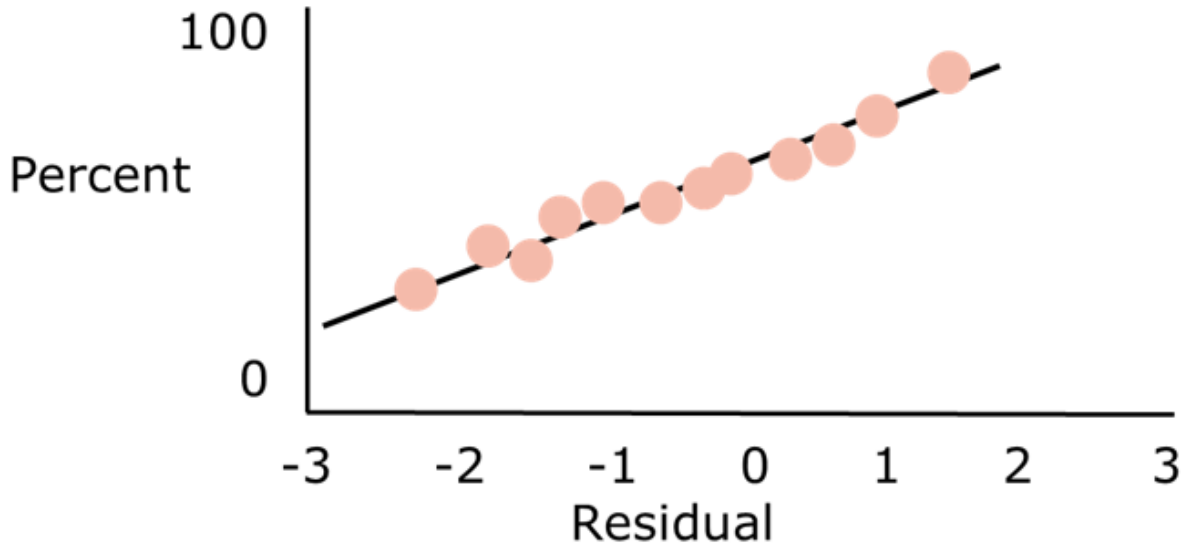


Checking for Normality

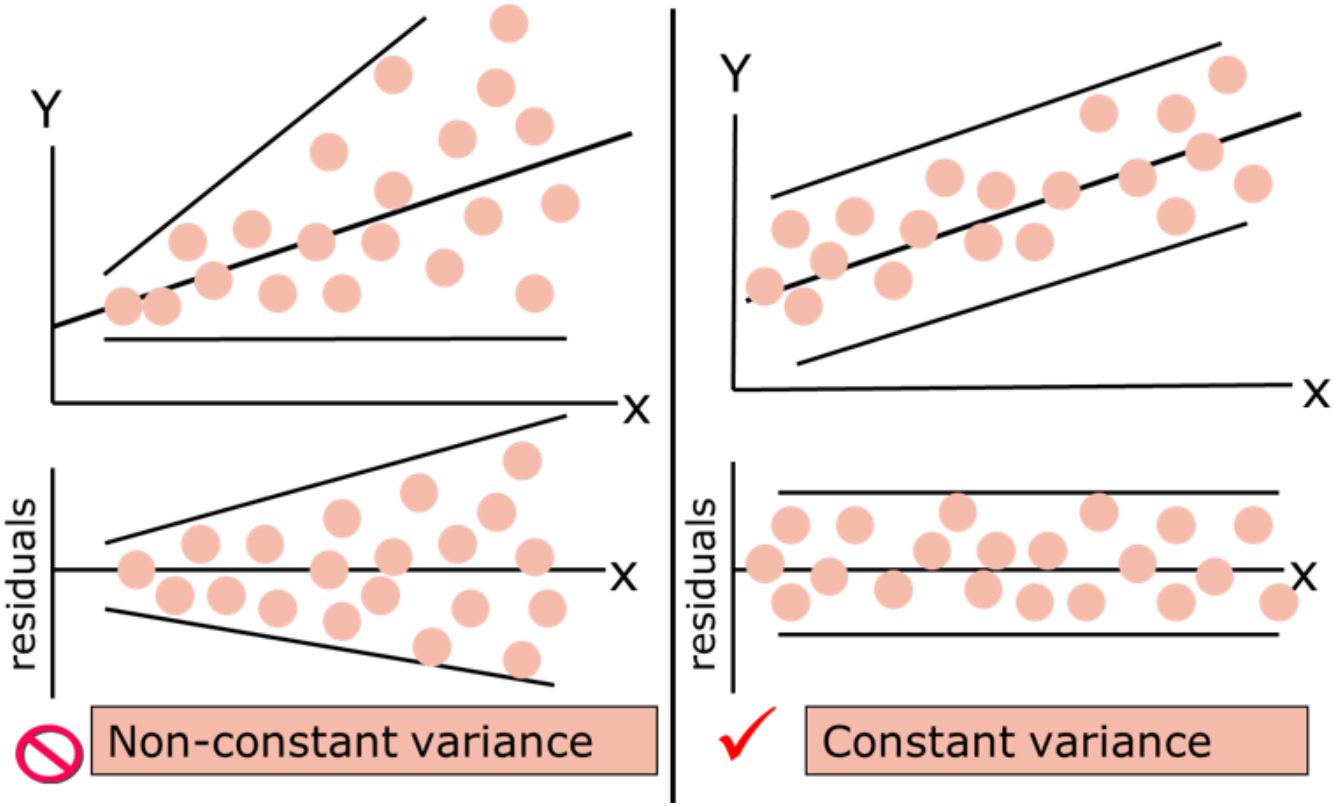
- Examine the stem-and-leaf display of the residuals.
- Examine the boxplot of the residuals.
- Examine the histogram of the residuals.
- Construct a normal probability plot of the residuals.

Residual Analysis for Normality

- When using a normal probability plot, normal errors will approximately display in a straight line.



Residual Analysis for Equal Variance

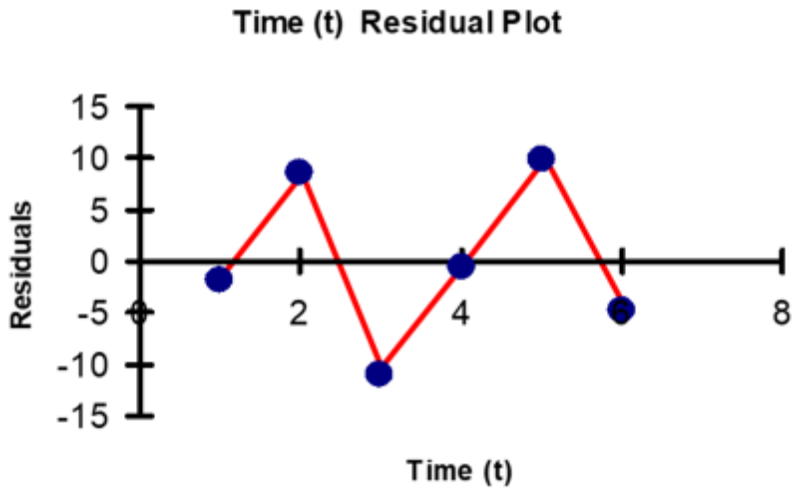


Measuring Autocorrelation: The Durbin-Watson Statistic

- Used when data are **collected over time** to detect if autocorrelation is present.
- **Autocorrelation** exists if residuals in one time period are related (dependent) to residuals in another period.

Positive and Negative Autocorrelations

- Positive autocorrelation, which occurs much more commonly in regression analysis, arises when there are clusters of residuals with the same sign.
  - e.g. Here, residuals show a cyclical pattern, not random. Cyclical patterns are a sign of positive autocorrelation.



- In negative autocorrelation, residuals tend to alternate in sign.

The Durbin-Watson Statistic (1 of 2)

- The **Durbin-Watson statistic** is used to test for autocorrelation.

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

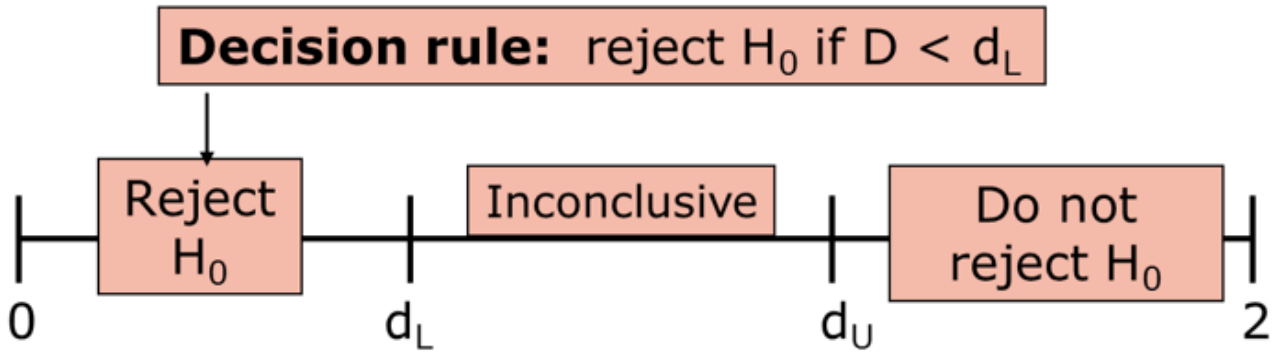
H_0 : positive autocorrelation does not exist
H_1 : positive autocorrelation is present

The Durbin-Watson Statistic (2 of 2)

- The possible range is $0 \leq D \leq 4$.
- D should be close to 2 if H_0 is true.
- D less than 2 may signal positive autocorrelation.
- D greater than 2 may signal negative autocorrelation.

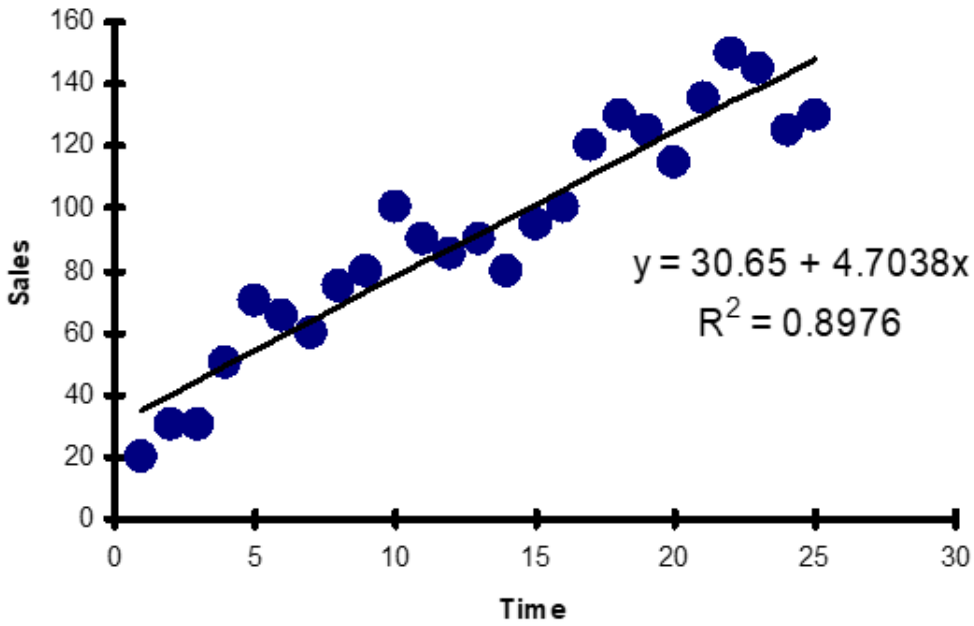
Testing for Positive Autocorrelation (1 of 4)

- Calculate the Durbin-Watson test statistic $= D$
  (The Durbin-Watson Statistic can be found using Excel)
- Find the values d_L and d_U from the Durbin-Watson table.
  (for sample size n and number of independent variables k)



Testing for Positive Autocorrelation (2 of 4)

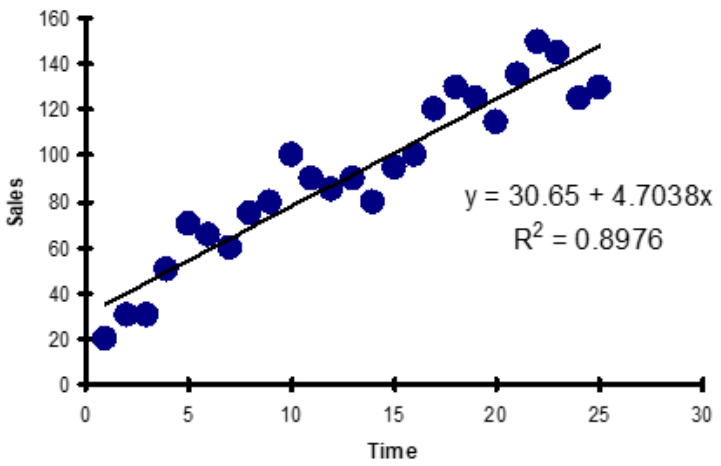
- Suppose we have the following time series data. Is there a positive autocorrelation?



Testing for Positive Autocorrelation (3 of 4)

- Example with $n = 25$.

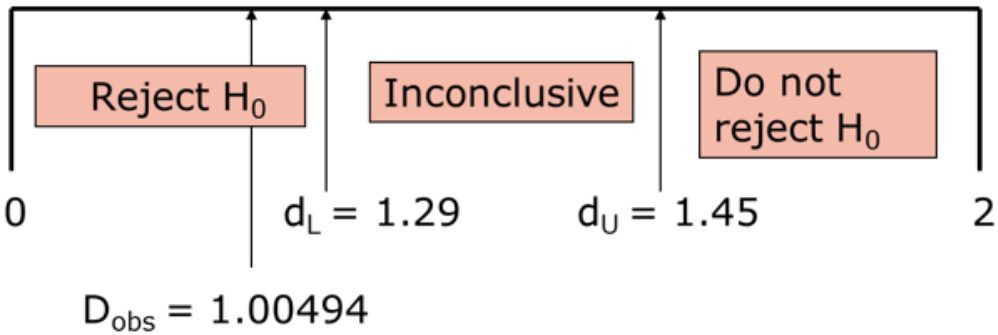
Durbin-Watson Calculations	
Sum of squared difference of residuals	3296.18
Sum of squared residuals	3279.98
Durbin-Watson Statistic	**1.00494**



$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$

Testing for Positive Autocorrelation (4 of 4)

- Here, $n = 25$ and there is $k = 1$ one independent variable.
- Using the Durbin-Watson table, $d_L = 1.29$ and $d_U = 1.45$.
- Decision: reject H_0 since $D = 1.00494 < d_L$.
- Conclude that significant positive autocorrelation exists.
- Therefore the linear model is not the appropriate model to forecast sales.



Inferences About the Slope

- The standard error of the regression slope coefficient (b_1) is estimated by:

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

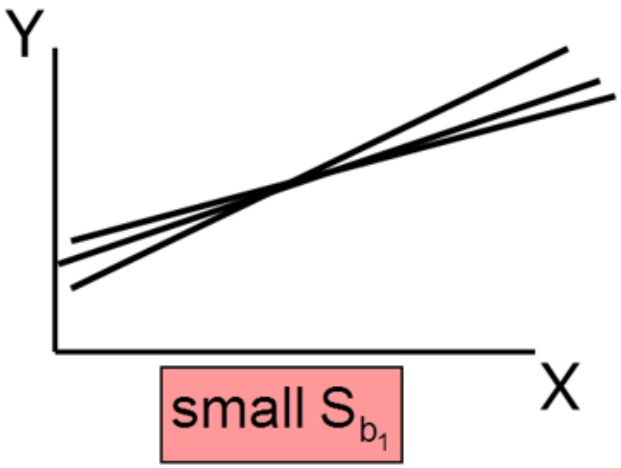
where

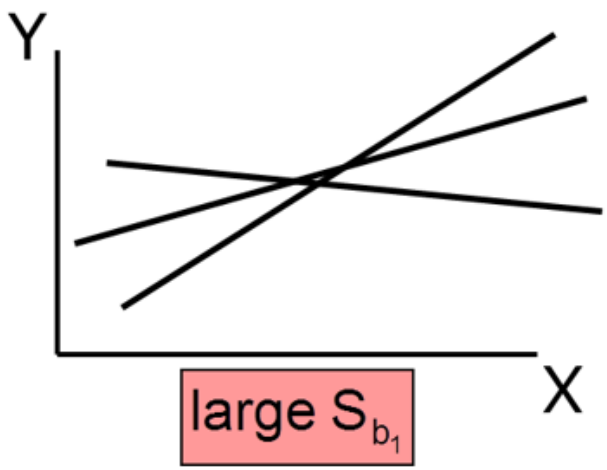
S_{b_1} = Estimate of the standard error of the slope

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

Comparing Standard Errors of the Slope

- S_{b_1} is a measure of the variation in the slope of regression lines from different possible samples.





t Test for the Slope

- **t test for a population slope**:
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses:
  - $H_0 : \beta_1 = 0$ (no linear relationship)
  - $H_1 : \beta_1 \neq 0$ (linear relationship does exist)
- Test statistic with d.f. $= n - 2$

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

where

b_1 = regression slope coefficient
β_1 = hypothesized slope
S_{b_1} = standard error of the slope

Example: t Test for the Slope (1 of 4)

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated regression equation:

house price $= 98.25 + 0.1098$ (sq.ft.)

The slope of this model is 0.1098.

Is there a relationship between the square footage of the house and its sales price?

Example: t Test for the Slope (2 of 4)

From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

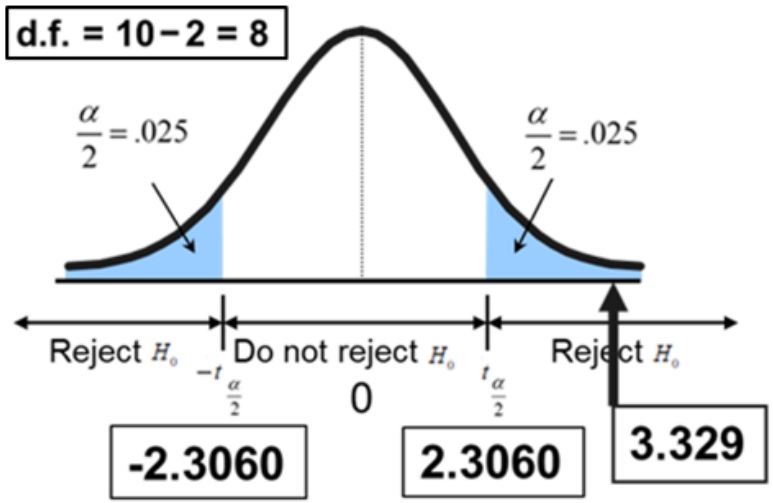
b_1

S_{b_1}

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Example: t Test for the Slope (3 of 4)

- Test statistic: $t_{STAT} = 3.329$



- Decision: Reject H_0
- Conclusion: There is sufficient evidence that square footage affects house price.

Example: t Test for the Slope (4 of 4)

From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

p-value

t Test for a Correlation Coefficient

- Hypotheses
  - $H_0 : \rho = 0$ (no correlation between X and Y)
  - $H_1 : \rho \neq 0$ (correlation exists)
- Test statistic with d.f. $= n - 2$

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

where

$$r = +\sqrt{r^2} \text{ if } b_1 > 0$$
$$r = -\sqrt{r^2} \text{ if } b_1 < 0$$

Example: t Test for a Correlation Coefficient (1 of 2)

Is there evidence of a linear relationship between square feet and house price at the 0.05 level of significance?

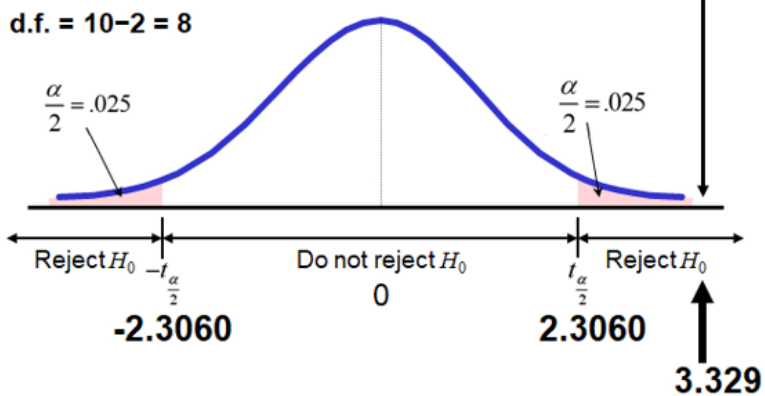
- $H_0 : \rho = 0$ (no correlation between X and Y)
- $H_1 : \rho \neq 0$ (correlation exists)
- $\alpha = 0.05$, d.f. $= 10$ -2 $= 8$
- Test statistic:

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.762 - 0}{\sqrt{\frac{1-0.762^2}{10-2}}} = 3.329$$

Example: t Test for a Correlation Coefficient (2 of 2)

- Decision: Reject H_0
- Conclusion: There is evidence of a linear association at the 5% level of significance.

$$t_{STAT} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.762-0}{\sqrt{\frac{1-.762^2}{10-2}}} = 3.329$$



Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression.
- Not knowing how to evaluate the assumptions of least-squares regression.
- Not knowing the alternatives to least-squares regression if a particular assumption is violated.
- Using a regression model without knowledge of the subject matter.
- Extrapolating outside the relevant range.
- Concluding that a significant relationship identified always reflects a cause-and-effect relationship.

Strategies for Avoiding the Pitfalls of Regression (1 of 2)

- Start with a scatter plot of X vs. Y to observe possible relationship.
- Perform residual analysis to check the assumptions.
  - Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity.
  - Use a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to uncover possible non-normality.

Strategies for Avoiding the Pitfalls of Regression (2 of 2)

- If there is violation of any assumption, use alternative methods or models.
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients.
- Refrain from making predictions or forecasts outside the relevant range.
- Remember that the relationships identified in observational studies may or may not be due to cause-and-effect relationships.